Critical Success Factors for A/B Testing in Online Marketing

Literature review

Bachelor's Thesis

Elja Lonkila

Information and Service Management

Fall 2019

Author Elja Lonkila

Title of thesis Critical Success Factors for A/B Testing in Online Marketing

Degree Bachelor's degree

Degree programme Information and Service Management

Thesis advisor(s) Xun Zhou

Year of approval 2019 **Number of pages** 42 **Language** English

Abstract

The increase of data-driven decision making in digital-facing organisations in the 2010's has brought methodologies such as A/B testing into the toolbox of online marketers. A/B testing in particular has become an essential part of the design process for advertisements, websites, and any user-facing interfaces.

This thesis aims to form a critical appraisal of A/B testing as a method by conducting a systematic literature review on how much the topic has been studied before. In addition, this thesis identifies the common pitfalls seen in implementation of A/B tests. The motivation to form a critical look into the subject rises from the rapid growth of the popularity of A/B testing. As the amount of companies utilising the methodology rises, it is important to review the topic to identify current best practices and possible deficiencies in research.

Keywords A/B testing, Online Controlled Experiment, Marketing

1. Introduction

1.1 The growing importance of data-driven decision making

In a modern digital-facing organization, data-driven decision making has become increasingly popular. A/B testing technologies have allowed scientific ways to design the layout of websites, advertisements, and product features and interfaces. Also known as online controlled experiments, split tests, and bucket tests, A/B tests have skyrocketed in popularity in the 2010's.

A/B test is essentially a randomized experiment between two variants of a website, advertisement, app, or any other product with a digital interface to find out which variant performs better. In practice, the users being tested are randomly split into two groups, who will see a different variant of the product that is tested. The users' interaction is measured, and a superior variant can be identified (Kohavi & Longbotham, 2017). This seemingly straight-forward method has changed the online marketing landscape substantially in the 2010's. The method has become a pervasive tool in marketing, and a critical part of doing business online after its beginning from the experiments of a handful of tech companies (Siroker & Koomer, 2013).

The popularity of A/B testing nowadays has brought the methodology to coffee room discussions. You might hear someone say "I A/B tested my email campaign by sending the same message with two different titles to 100 people". However, A/B testing is not applicable to all organizations and all decision making processes. In its core, A/B testing is an application of statistical hypothesis testing, where the result is deemed by statistical significance, which is why making decisions based on the results must be backed by statistically significant evidence leading to the rejection of the null hypothesis. Therefore,

in the ideal situation, the experimenters need to have an understanding of the basic statistical concepts related to the probability testing.

Successful brands such as Netflix and Amazon have built an experimentation culture, where the approach to decision making is data driven, and thousands of A/B tests are run constantly. Experimentation is utilised not only in marketing, but also in supply chains, brand value propositions, and more (Accenture 2018).

The method offers substantial advantages to businesses, when applied correctly, but there are also many pitfalls to be wary of (Kohavi & Longbotham, 2007). In this literature review, a critical appraisal of A/B testing as a method will be formed to highlight critical success factors when conducting online experiments.

1.2 History and background of A/B testing

As reported by Rossi et al. (2003) and many others, the beginning of the narrative of A/B testing dates back to the 16th century , when a captain of a British ship experimented with his crew's food rations after he had noticed a lack of scurvy among sailors based in the Mediterranean region. A treatment group was formed, who received vitamin-C rich limes in their rations, while a control group continued with the same diet as before. Lack of scurvy was observed in the treatment group, and soon all sailors would receive limes in their rations.

In an online marketing context, the methodology was adopted in to commercial use much later. Kohavi et al. (2017) reported that the first online experiments date back to the late 1990s. A monetary value to experimenting was discovered later when a blog post from an Amazon engineer Greg Linden from 2006 talked about a "fun project" he had made for shopping cart recommendations for the Amazon online store. This project was an online controlled experiment (also known as an A/B test) and it turned out to have an effect to the

revenue of the online store. Soon after, academia started to notice the phenomenon. The first notable academic paper on A/B testing in online marketing context is from 2007, where Kohavi, Henne, and Sommerfield, engineers from Microsoft, state that conducting controlled experiments online forms a better representation of how a business's customer base reacts to changes on a website, than a "HiPPO" (Highest Paid Person's Opinion). In 2009, Kohavi et al. follow up on their studies on online controlled experiments with an extensive library of real-life cases from Microsoft, where A/B testing was utilised and found effective. Early examples include experimenting with features of MSN and Microsoft Office.

Soon other engineers followed suite in ramping up online experiments in their own teams. Kohavi and Thomke (2017) find an early example in their Harvard Business Review article where they state that a Bing.com engineer had experimented with a single small controlled experiment in 2012 that led to a 12% increase in revenue. In the coming years other companies would follow, and today companies such as Google, Facebook, Microsoft, and Amazon run over 10000 experiments annually. Kohavi and Thomke (2017) state that Bing has been able to increase revenue from 10% to 25% annually by utilising controlled experiments. Although the experiments started from companies with digital roots, the benefits to A/B testing have been identified in non digital-native companies as well. Nowadays A/B testing is utilised in a wide array of industries, such as retail, airlines, and transportation.

In a survey on conversion optimisation published in 2018, conducted by ConversionXL and VWO, A/B testing was used in 98% of the responders' companies. The number was 90% in 2017, and 70% in 2016, indicating a recent spike in popularity for the method. For a digital facing company in 2019, A/B testing seems to be quintessential for success.

1.3 Research objective and research questions

The objective of this thesis is to form a comprehensive critical appraisal of A/B testing applied in online marketing and give the reader of this thesis an understanding to what the critical success factors for A/B testing are in online marketing. The thesis aims to analyse the topic from both managerial and a theoretical standpoint, and identify critical success factors for A/B testing that affect the success of the experiment. An appraisal of the method's credibility will be evaluated by conducting a systematic review on previous research and use cases. The reader of this thesis will be provided with an overview of A/B testing as a method, the practical applications of the method, and challenges managers face when conducting these controlled experiments.

The topic is young in academia, and a relatively low amount of articles have been published of it in the past considering it's large popularity in business applications. The literature related to the topic hasn't been reviewed systematically before, which in part motivated the making of this thesis. The motivation to conduct this thesis arises from these factors, as it is beneficial to find out if some areas of the topic has been overlooked in academic literature, or other discrepancies in studies can be identified.

The thesis does not go into extensive detail on the statistical and mathematical theory behind A/B testing, but aims to give an overview of the subject and familiarize the reader with the theoretical concepts related to the topic. Key concepts are explained in short in chapter 2.

The specific research questions the thesis aims to answer are:

1) How much has A/B testing been studied before?
2) What are the critical success factors for A/B testing in online marketing?

To answer the first question, a broad systematic review of scientific articles and conference papers on the topic will be conducted and the findings reported. In addition to volume of research, information will be gathered on what methods have been used in previous studies, what trends can be identified, who are the most established authors in the field, and which subject areas have the most relevance to the topic.

To answer the second research question a critical review will be formed of the topic using selected literature as references. To find relevant literature, keyword searches, and backward and forward searches from selected articles are utilised. More detailed explanation to criteria for source inclusion is found in chapter 3.3.

1.4 Structure of the thesis

The thesis is divided into five chapters. The first chapter contains a brief overview of the topic and its history and background. In the first chapter, motivation to the study, the research objective and questions are specified.

The second chapter explains the key concepts and terminology related to A/B testing briefly. The reader is introduced to the basic statistical concepts behind the methodology, and how they relate to A/B testing.

The third chapter contains details of the methodology behind the literature review, specifying what databases, keywords, and techniques will be used in the database search. Finally, the third chapter goes into detail on the criteria for source inclusion used in this thesis.

In the fourth chapter, the results of the search will be presented to provide the reader with an overview on the topic. After the first section of the fourth chapter the reader should have

an understanding on how much the topic has been studied before, and what are the current trends in research. The reader will be then introduced to the second research question and provided a detailed view into the critical success factors for A/B testing in performance marketing. The fourth chapter is divided into subsections, which cover the main findings from the literature review on the subject.

Finally, the fifth chapter wraps up the thesis with discussion, key take-aways, and conclusions on the critical success factors for A/B testing in performance marketing. The findings will be analysed from both managerial and theoretical standpoints, and limitations to the research and topics for further research are identified. The final chapter contains references used in the study and any appendices, such as tables or charts.

2. Key concepts and terminology

This chapter aims to make the key concepts and terminology related to A/B testing clear to the reader of this thesis. As the research areas vary in literature on A/B testing, the terminology varies as well and many concepts appear under different synonyms in different articles. To align the concepts within the context of this study, a textbook by Siroker and Koomen (2013) on A/B testing in online marketing, and a statistics textbook by Ross (2005) is used as reference.

Hypothesis testing, as mentioned in chapter 1.1, is the theoretical basis for A/B testing. When making decisions based on measured data of a real world event/population, hypothesis testing is used to see how likely it is the measurement reflects the real population. Hypothesis testing is conducted by defining the *null hypothesis*, or a normal case you are trying to disprove. An *alternative hypothesis* is formed based on what the experiment is trying to prove. From the measured sample the experimenter can then deduce the likeliness of the alternative hypothesis holding, i.e. rejecting the null hypothesis, by calculating test measures. The possible outcomes are:

1) If the result is *very unlikely* under the null hypothesis, conclude that the alternative hypothesis is true.
2) If the result is consistent with the null hypothesis, conclude that the null hypothesis is true.

The reader may ask: What if the very unlikely happens? This comes down to *experiment design.* Before conducting the experiment, the experimenter must choose the probability he/she is comfortable with that the value of a parameter from the population falls within the level of values concluded from the experiment. The probability is called the *confidence level.* Most common confidence level used in business applications is 95% (Siroker and Koomer, 2013).

In the (unlikely) case that the experimenter concludes something about a population that is not in fact true, an error occurs. There are two kinds of errors in statistical hypothesis testing, *Type I errors (False positives)* or *Type II errors (False negatives)*.

	Do not conclude H_1	Reject H_0, conclude H_1
H_0 is true	True negative	Type I error
H_1 is true	Type II error	True positive

All possible outcomes of hypothesis testing (Ross, 2005).

The probability for these errors can be determined. The probability for Type I errors is called *statistical significance*. Similar to confidence level, the experimenter chooses the significance level he/she is comfortable with. Significance level is the complement of confidence level, and usually set at 5% in business applications ($1 - 95\%$). To test the significance of the test, experimenters calculate a *p-value* from the sample to tell the probability of getting an "extreme" result under the null hypothesis. If the probability is below the critical value for statistical significance, the null hypothesis can be rejected.

The *power* of a test is the probability to correctly reject the null hypothesis, assuming it is false in the population. Essentially, the power of a test is the complement of the probability of committing a Type II error, i.e. failing to conclude with the alternative hypothesis when the null hypothesis is false.

Decreasing the significance level reduces the amount of Type I errors, but increases the amount of Type II errors, therefore making experiment design a trade off between significance level and power. The only way to influence the amount of Type II errors by experiment design is to increase the sample size, which reduces the variances.

Due to the uncertainty that is present in experimenting, hypothesis testing is considered a decision process, not a proof. Essentially the only thing hypothesis testing does is ask if the result is inconsistent enough with the null hypothesis to question or reject it.

In A/B testing, the experiments are usually *tests for proportions*. The experiments are set to measure a single data point that can be quantifiable under success or failure. This data point is called a *variant* or a *factor.* The metric used for hypothesis testing is then the number of successes divided by the number of trials, i.e. the *proportion*. For example, a typical A/B test would measure the proportion of users that clicked on a button that was blue, and compare the result to the proportion of users that clicked on a button that was red. Based on these proportions, an estimation of the likelihood for the whole user base (population) clicking a red button can be calculated. The experimenters essentially ask the question: "If a specific change is introduced, will it improve my key metrics?" (Kohavi et al. 2017)

Overall Evaluation Criterion (OEC) is a quantifiable measure of success of the test. A good OEC is a single, short-term focused metric (Kohavi et al., 2009). Examples of an OEC used in A/B testing include revenue, profit, time spent on page, or other quantifiable metrics that are believed to have an effect on the long term.

As reported by Accenture, the traditional Overall Evaluation Criterion (OEC) in A/B testing is usually pictured to be a myopic measure, such as click-through rate. However, recently a shift towards more holistic measurements has occurred, with measurements such as Customer Lifetime Value and Net Promoter Score being used as OEC for online experiments.

3. Methodology

The research will be conducted as a literature review on scientific journal articles and textbooks. In addition conference proceedings, business magazines, and research publications will be used to support any findings. From these resources a critical appraisal on the topic will be created.

To answer the first research question, a focused literature review is conducted to form a picture on how much the topic has been studied before, what methods have been used, and what trends can be identified. This search will be conducted systematically from the Scopus and Web of Science databases using selected relevant keywords. The literature from the keyword search will be categorized by year of publication, document type, and field of study.

Systematic review as a method is advantageous in the context of this thesis, as the topic is relatively young, and the amount of research conducted is moderate. Systematic reviews aim to reduce bias in the results, and create an exhaustive outlook on the subject. Although at times time-consuming, systematic review is considered to be an accurate and unbiased method for evaluating extensive literatures with the advantage of considering a larger amount of publications than a traditional narrative review would (Mulrow 1994).

This thesis does not go into extensive detail on the meta-analysis of the articles found in the systematic review. To evaluate qualitative studies, a detailed look into each article would be required to uncover data from the text. For the purpose of this study, the main data points for articles will be number of results, year of publication, document type, and field of study. Avoiding subjective and qualitative measures such as authors ability to answer research question reduces error and bias in the review.

Outside of the systematic review, databases Scopus, Google Scholar, and Web of Science will be utilised in a critical fashion to conduct backward and forward keyword searches to find relevant literature in order to identify reputable sources to use as references to the research. The criteria for source inclusion will be explained in detail in chapter 3.3.

3.1 Databases and keywords used in database review

To form an understanding on how much the topic has been studied in the past, a systematic review will be conducted using selected keywords. The aim is to identify trends and possible deficiencies in research.

Scopus and Web of Science (WoS) will be used as the databases for this search. The databases offer advantages for this kind of systematic keyword review due to their heavy curation and good reputation among the academic community. A more comprehensive search could be done using databases such as Google Scholar, but for the purpose of this study, a narrow scope is preferred.

The keywords used are chosen based on the most common spelling suites and synonyms of the methodology. Using boolean operators, the keywords are combined to form a more complete picture of the research conducted. For example, *"A/B Test*"* returns results for both *"A/B testing"* and *"A/B test"*, which often appear as keywords in different articles. Using search strings such as *"A/B test* OR "online controlled experiment"* broadens the search to include synonyms and different terminology of the method.

In the context of this study, focusing on marketing and business literature is important. The spelling suits of *"A/B testing"* vary by industry. For example, *"ab testing"* returns articles related mostly to medicine, immunology, and microbiology. A more relevant field to the study uses keyword *"online controlled experiment"* more often, and searches using this keyword return results from the fields of computer science, business and administration,

and mathematics. Other forms of spelling such as *"split testing"* and *"bucket testing"* have appeared in academic literature as well, although not very prominently and often as secondary keywords to studies referring to A/B testing.

Majority of academic research on A/B testing is from the field of computer science. In the context of this research, studies from the field of marketing and business administration are especially relevant.

Exploratory keyword searches proved the most relevant keywords for the purpose of this study to be *"A/B testing"*, and *"online controlled experiment"*. The systematic keyword review is completed with the search string *"A/B testing" OR "online controlled experiment"*. To tie the topic into the field of marketing and business, a systematic review is conducted on the search string *"A/B testing" AND "marketing"*.

3.2 Other relevant literature used in the research

To answer the second research question, a critical review is conducted. The critical review is conducted from various sources found from database searches and included based on criteria introduced in chapter 3.3.

As the research conducted on the topic is often quite recent, there isn't an extensive amount of scientific articles on A/B testing in marketing literature. Most publications used in the research are from computer science publications, focusing on the theoretical viewpoint of A/B testing. To support the literature review, articles from credible marketing journals and websites, along with reports from digital consultancies are used in the thesis. The statistical concepts are explained to the reader using statistics text books as reference.

3.3 Criteria for source inclusion

The credibility of the sources used in this thesis are appraised using the following criteria commonly used for source evaluation: Accuracy, Authority, Objectivity, Currency, and Coverage (AAOCC). Most databases for academic literature are already filtered quite strictly in terms of these criteria. Appraisal of the sources used gains importance when looking outside of academic literature, in the context of this thesis, reports and web articles.

4. Results & Findings

In this section, the findings from the systematic and critical reviews are presented and analysed to answer the research questions presented in chapter 1.3. First, the findings from the systematic review are presented in chapter 4.1. Analysis is conducted based on the results to form an answer to the first research question.

From chapter 4.2 onwards, the thesis deals with the second research question, and is based on a critical review. The reader is introduced to successful cases of A/B testing and the advantages of using the methodology correctly to provide context on why the method has gained substantial popularity among marketers globally.

As mentioned in the introduction, A/B testing isn't all that straight-forward as the successful cases may sound. To present the reader with an understanding of what can go wrong when conducting experiments, commonly mentioned pitfalls in the execution of A/B testing are gathered and presented in chapter 4.3. The reader will get a basic understanding of what are the factors that often affect the success of an experiment. In chapter 4.4, a more detailed look into metric design and interpretation is formed. Metric design is an especially relevant topic in marketing concept, as it is often mentioned in literature as a critical factor for the success of the experiment.

In chapter 4.5, a brief review of current trends and future outlooks for A/B testing is conducted to provide the reader with a more complete picture of the method, and on what may change in the future when human error is being mitigated with artificial intelligence.

4.1 Results of systematic keyword review

In this chapter, the results of the systematic review are presented and analysed. Each search string for systematic review is analysed using a standardized structure, where the following statistics are evaluated:

1) Publications by year
2) Document types
3) Research areas

A brief analysis is conducted from the findings of each keywords and possible trends in research identified.

4.1.1 Results of the first review

For the first search, search string *"A/B testing" OR "online controlled experiment"* was used. Exploratory analysis of the keyword-specific results had shown most relevant results in terms of citations and research area for the keywords *"A/B testing"* and *"online controlled experiment"*.

Figure 1 presents the amount of publications that have either *"A/B testing"* or *"online controlled experiment"* as a keyword in WoS and Scopus databases from the year 2006 to 2019. Before 2006, a handful of articles could be found with keyword "*A/B testing",* but none relevant to this thesis. The total number of articles was 450 in Scopus and 214 in WoS. First publications under keyword "*online controlled experiment"* begin from the year 2009.

The research conducted on the topic is quite recent. As stated in chapter 1.2, A/B testing started gaining popularity in the early 2010's. Practitioners from companies like Google and Microsoft were able to show significant results using online experiments, which awakened the interest of academics. As the figure shows, the number of publications

rapidly grew from 2010 to 2015. The amount of publications has been on the rise since, signaling the popularity of A/B testing. The data was accessed on 19.10.2019, explaining the slight drop in the amount of publications in 2019.

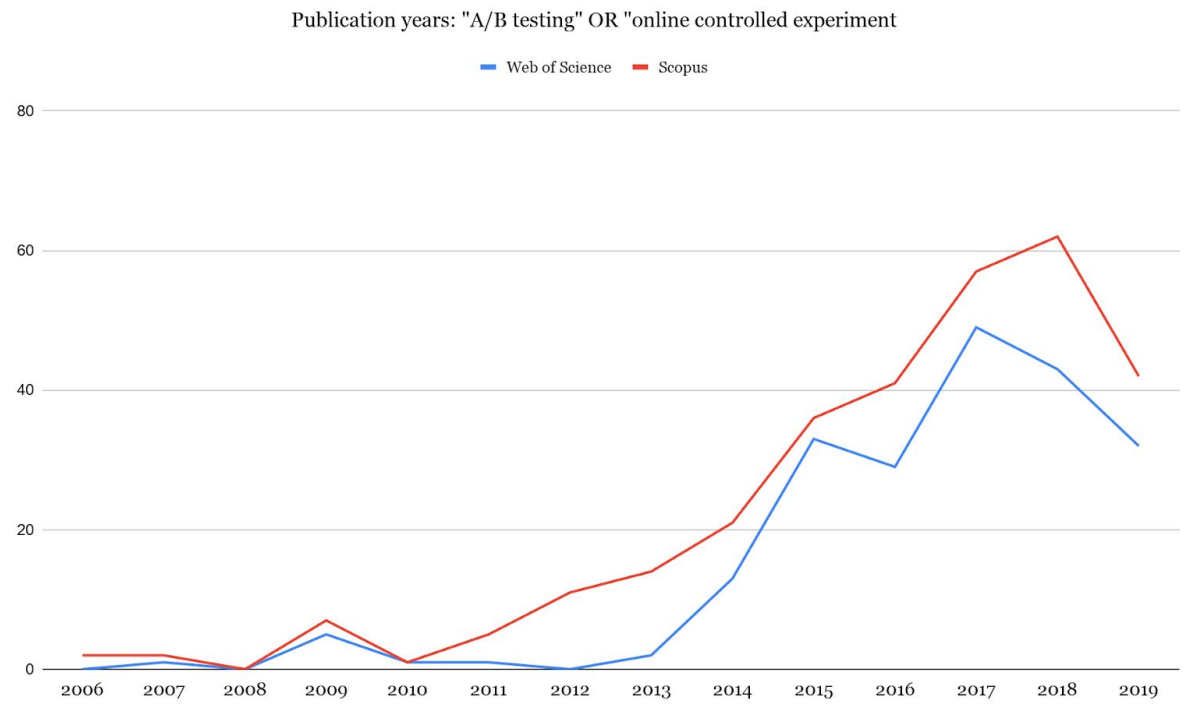

Figure 1. Number of publications from Web of Science and Scopus databases from 2006 to 2019 for the search string "A/B testing" OR "online controlled experiment". Data accessed on 19.10.2019.

Tables 1 and 2 in the appendix present the subject areas related to the search string. The majority of literature is from the field of computer science, with over 55% of articles in Scopus and 80% in WoS classified under the subject area. Engineering, mathematics, decision sciences and business are also found to be significant subject areas for the keywords. It is worth noting that the classifications for subject areas are done differently in

Scopus and WoS. WoS allows an article to hold multiple classifications for research area, explaining why the sum of the percentages in table 2 is over 100.

Tables 3 and 4 in the appendix present the document types for the search string. Majority of the literature found has been published as a conference paper, with 74.3% of WoS articles and 50.22% of Scopus articles classified under the document type. The second most prevalent document type in the search is article, with 27.1% of WoS and 12.89% of Scopus literature falling under the category. Articles generally go through a stricter and longer review process than conference papers, which is why a trending topic such as A/B testing hasn't seen a significant amount of articles published. Conference papers often allow for a quicker publication process, as the review cycles for the papers are usually more streamlined and under fixed schedules.

The systematic review for the search string *"A/B testing" OR "online controlled experiment"* highlights the relevancy and currency of the topic. A growing amount of research on the subject with the majority of literature published in conferences signal that the topic has a significant foothold in academia and practice.

4.1.2 Results of the second review

The second search string used in the systematic review is *"A/B testing" AND marketing.* The reasoning is to tie the computer-science heavy topic to the context of this thesis, online marketing. As the systematic review shows, the amount of literature on the topic in a marketing context isn't very extensive, with 26 results found in the Scopus database for the search string and 28 results found in WoS.

Figure 2 presents the amount of publications that have the keyword *"A/B testing"* and *"marketing"* from the year 2006 to 2019. Using the "AND" operator cuts down the number of publications found and results in a more concentrated sample of literature. The data was

accessed on 19.10.2019, explaining the dip in the figure for that year. Overall, the total number of articles is very low for this search string. It is apparent that A/B testing hasn't been excessively studied in marketing context.

Although minor in volume, the rise in publications in the early 2010's is evident. Literature on A/B testing with the keyword "*marketing"* peaked in the amount of publications in the years 2016 & 2017.

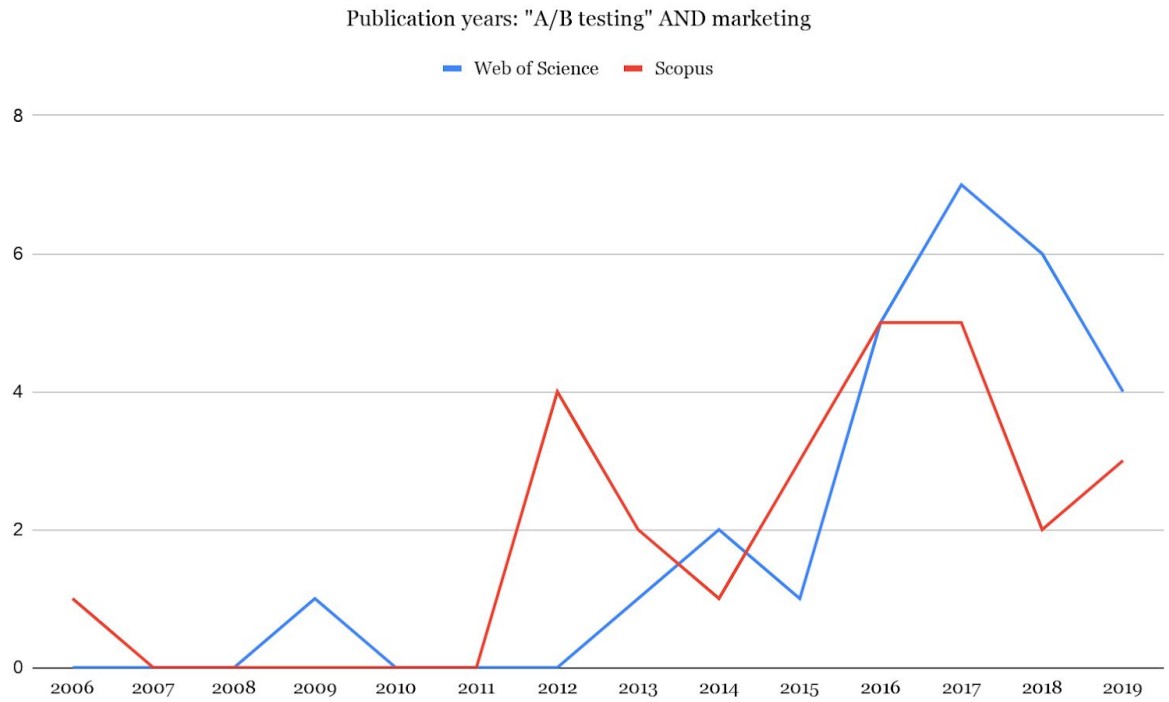

Figure 2. Number of publications from Web of Science and Scopus databases from 2006 to 2019 for the search string "A/B testing" AND marketing. Data accessed on 19.10.2019.

Tables 5 and 6 in the appendix present the research areas for the search string. Computer science is the dominant research area, with 76.92% of Scopus and 82.14% of WoS results classified under it. Second most prominent research area is Business, with 15.38% of Scopus and 67.86% of WoS results in the category. Comparing the results to tables 1 and 2

confirms the assumption made in chapter 2.1 that the search string ties the results closer to business literature.

Tables 7 and 8 in the appendix present the document types for the search string. As with tables 3 and 4, the majority of literature in the databases is classified as conference papers with 65.38% of documents in Scopus and 71.43% of documents in Web of Science falling in the category.

4.2 Advantages of correctly implemented experiments

Online controlled experiments are considered essential to a digital-facing company's marketing toolbox, and described as an "indispensable tool" for both startups and large enterprises (Kohavi et al., 2017). The credibility of the method is based on its foothold in the operations of the majority of companies working in the digital space.

Two example cases of successfully implemented experiments are presented in this chapter to provide the reader with an understanding of why experimenting is worthwhile, and how it can contribute towards real-life business goals. The examples chosen are often referenced in literature.

Siroker and Koomer (2013) present an example of successful implementation of A/B testing from former US president Barack Obama's election campaign in 2012, where A/B testing the donation button and landing page images on Obama's campaign website led to an increase of 57 million dollars in donations. Siroker argues that behind Obama's election success was the media team who were "willing to test everything and listen to the data even when it surprised them the most."

Kohavi et al. (2017) present an example on an experiment that ran on Microsoft's search engine website Bing.com. The team hypothesised that adding links to a site's subcategories

in search engine advertisements would increase traffic. Figure 3 shows what the experiment looked like in practice.

Figure 3. Online controlled experiment worth tens of millions of dollars (Kohavi et al., 2017)

The OEC of the test was increasing average revenue without a decrease in other engagement metrics. The results showed an increase in revenue, but a decrease in page load time and user metrics. This simple feature increased ad revenue by "tens of millions of dollars" yearly, while keeping the impact on other metrics neutral (Kohavi et al., 2017). The return on investment for such an experiment is extremely high, as the only cost is essentially labor.

The advantages of optimising marketing efforts by data-driven methods are apparent from cases such as Obama's and Bing's. Today, any digital-facing company can effortlessly access A/B testing tools and start experimenting on their marketing with relatively low investment. Multiple companies, such as LeanPlum, Apptimize, Optimizely, Taplytics offer testing solutions for companies that don't have resources to set up their own in-house systems similar to Google's and Microsoft's (Dmitriev et al., 2017).

4.3 Common pitfalls seen in implementation

The length of the experiment is often identified as a critical factor for the success of an experiment in academic literature. Dmitriev et al. (2016) argue that short term changes to a testable metric often don't translate into long-term results. Lu and Liu (2014) call this phenomenon the "novelty effect". For example, an increase in revenue short-term does not indicate a higher customer lifetime value. Hohnhold et al. (2015) state that a remedy to this is to conduct long-term experiments. The experimenter is presented to a wide array of possible pitfalls when conducting these long-term experiments. As pointed out by Keser (2019), cookie churn becomes a significant issue for testers when conducting experiments on the long term, meaning the users clear their browser cookies mid-experiment resulting in polluted data.

Another issue specific to long-term A/B tests is user learning as presented by Hohnhold et al. (2015). User learning is used to describe the scenario where positive outcomes reinforce behavior leading to it. The effect is similar to psychologist Thorndike's law of effect, where it is stated that actions leading to positive outcomes are likely to be repeated. In an online marketing context, such a scenario could be for example the change of the likeliness of a user to click on an ad after been exposed to multiple similar ads. Dimitriev et al. (2016) state that in lengthy online controlled experiments, this may lead to bias in the results, as learning is more likely to happen and difficult to measure. They also identify the tendency for survivorship and selection biases in long-term experiments. Survivorship bias in online experiments essentially means that the users are abandoning the experiment at different rates, resulting in biased results. Selection bias is used to describe a scenario where the test sample is already biased toward a certain segment of users. (Deng et al., 2016; Dmitriev et al., 2016).

A common mistake seen in online experiments is the multiple comparison problem (Esteller-Cucala et al., 2019). Often the experimenter wants to test slight changes in a

feature by introducing more variants. Introducing more variants has an effect on the significance of the results and quickly the likelihood of Type I errors increases. In addition, a larger sample size is needed to reach significant results, often meaning a lengthier test. Keser (2019) states that lengthier tests lead to polluted data, as people may delete their browser cookies and end up seeing a variation they weren't meant to see. Esteller-Cucala et al. (2019) conclude that there are significance-level corrections that can be applied to combat this problem. Keser (2019) also points out that modern tools often have built in features to negate the issue.

The type of change measured is also influential to the outcome of the experiment. Kohavi et al. (2014) present examples from Microsoft's experiments, and find that while small, incremental changes may have a significant impact on key metrics, the cases are rare, which is why organizations should also test larger changes in features as well. A practical guide by VWO, a widely used optimization platform, states that the experimenter should conduct website and visitor data analysis in order to find elements more likely to have an effect on the OEC.

"Twyman's law" states that scientific results that look "too good to be true" are usually wrong. In literature addressing A/B testing pitfalls, Twyman's law gets mentioned regularly. The experimenters often want to believe that their change of a variant resulted in a breakthrough. Kohavi et al. (2014) state that these kinds of breakthroughs are extremely rare and often the "too good to be true" results are caused by a bug or a Type II error. Blindly adopting unexpected results from a test can be described as an effect attributed to human behavior. Unfavorable results are analysed with more scrutiny to find an issue in the test, when favorable results are more likely to be accepted even if unexpected (Esteller-Cucala et al., 2019).

As reported by Bakshy and Eckles (2013), dependence in experiments is often a critical issue. Despite the experiment data consisting of multiple individual data points, statistical

inference is a concern. Ignoring this concern may lead to an increase in Type I errors. Kohavi et al. (2017) point out that Microsoft had identified hundreds of invalid experiments by being skeptical towards the quality of the data and formulas used. One common inaccuracy they mentioned was assuming the variables to be independent when in fact, they often aren't.

4.4 Metric design and interpretation

In A/B testing the success of the experiment is measured by the OEC, also known as key metric, or goal metric. Companies such as Amazon, Facebook, Google, eBay, and Netflix rely on online metrics as an indicator of the success of their business. When making business decisions based on these metrics, the metric development plays an important role. Aligning the metric to reflect the real world behavior of users requires a systematic, data-driven approach to metric design (Deng and Shi, 2016).

Metrics are used in contexts outside of A/B testing as well, for example in reporting, dashboard, analyses. However, in A/B testing the metrics are a scientific indication of how the implemented feature has influenced performance, with limited sensitivity to external factors (Dmitriev and Wu, 2016).

Deng and Shi (2016) introduce a framework for categorising different types of metrics. They classify metrics by type into three different categories. *Business report driven metrics* are designed based on the long-term goals of the business, such as Revenue per User, or Monthly Active Users. To measure long-term impact, the experiments usually need to be ran for a longer time than in a feature experiment. *Simple heuristic based metrics* are short-term, actionable metrics that indicate how the user interacts with different features. The third type, *User-behavior-driven* metrics measure user behavior to find out which emotions the user experiences under the experiment. This can be measured, for example, with frustration and satisfaction models. When implemented correctly, these kinds of

metrics can be applied in both long- and short-term experiments, and they aim to provide actionable insights on user behavior that can't be gathered from simple metrics such as click-through-rate.

Metrics are evaluated by their properties and qualities. Hauser and Katz (1998) find two of the most critical properties for a metric to be:

1) A change in the metric should affect the company's ability to reach it's long-term business goal.
2) Individual contributors (employees) should be able to impact this metric.

Dmitriev and Wu (2016) agree that these two properties are essential for a company to consider when designing metrics. They introduce several "meta-metrics" to help evaluate the quality of the metrics, keeping in mind the two abovementioned main properties. *Sensitivity* of a metric is a description of how much data is required for the metric to show a statistically significant change. Sensitivity is affected by the amount of data, the variance of the metric, and the size of the effect. Deng et al. (2013) state that increasing the sensitivity of a test leads to more precise assessment of the metric. They argue that sensitivity can be increased by using pre-experiment and offline data. Other "meta-metrics" introduced in Dmitriev and Wu's study include alignment with user value, and automation.

A lot can go wrong when interpreting metrics. Dmitriev et al. (2017) identify twelve common misinterpretations of metrics in online controlled experiments. From the research group's work on thousands of experiments at Microsoft, it became apparent that human misinterpretation of metric movement was often the fault of an experiment failing leading the experimenters with false data on how a change, if deployed would affect the metric set for the experiment. Ultimately, savviness in metric design and interpretation can save a company millions of dollars.

4.5 Future applications and automation in A/B testing

To mitigate human error in experimenting and drive efficiency, researchers have come up with mathematical models to provide automated solutions to A/B testing. Modern platforms such as Leanplum and Optimizely have already implemented features that automatically calculate statistical significance, and use business intelligence to provide "intelligent insights" about the test results.

Tamburelli and Margara (2014) provide an early framework for automating A/B testing in their paper "Towards automated A/B testing". Using genetic algorithms, a machine learning model can be implemented to automatically select the most efficient factors or variants in the experiment. Empirical research on this topic is still very scarce, but early prototypes have been experimented with. Cruz-Benito et al. (2017) report on their prototype for a machine learning model to generate predictive insights for experimenters, but don't conclude on the possibility of real-life industry applications for the model.

Fabijan et al. (2018) in their survey on the state of A/B testing find that automation capabilities still vary within companies conducting experiments. Many companies have their in-house solutions for conducting experiments, explaining the difference in the maturity for experimenting. The volume of research on automated A/B testing has increased recently. Although still moderate in volume, the amount of publications on machine learning in A/B testing context has been trending upwards since 2016.

5. Conclusions

In this thesis A/B testing in online marketing was analysed from an academic and a practical standpoint by conducting both systematic and critical reviews on the subject. This chapter collects the findings and provides reasoning and analysis behind the results.

The analysis on conclusions are divided into implications for academic research and practice. The implications to research presented are the writer's conclusions drawn from the results of the thesis and recommendations for further research. Implications for practice are highlighted to provide the reader with key points on what to conclude of the thesis and what to keep in mind when dealing with experiments in practice. Finally, limitations to the thesis and possible avenues for future research are identified.

5.1 Research implications

The research objective of this thesis was to form a comprehensive critical appraisal of A/B testing in online marketing context. The literature review revealed the topic to be relatively young in academia, and the popularity of the method growing rapidly in both real-life business applications and the amount of papers published on it.

Since the topic hasn't been analysed thoroughly in the academic spotlight, a critical appraisal was called for. Although many areas of A/B testing have been studied, and critical success factors identified, the research on the subject is no way exhaustive. The methodology is still young, and only in the latest years it has been adopted to use in the majority of the digital-facing companies. The best practices are still being formed and research continues to find remedies to challenges faced by the experimenters.

5.1.1 Implications for academic research

A great proportion of literature referenced in this thesis is conducted by Ron Kohavi, or other Microsoft engineers. Kohavi and his research group have written and influenced the most important academic papers on A/B testing if measured by citations. Kohavi and his team have done groundbreaking work in studying A/B testing, identifying challenges in the methodology, forming best practices, and inspiring other researchers. In the context of this thesis, the affiliations of the literature reviewed leave an asterisk to the breadth of the review.

Table 9 indicates documents by affiliation for search string *"A/B testing" OR "online controlled experiment"* in the Scopus database. The most prevalent affiliation by far is Microsoft, followed by LinkedIn and Yandex. Especially Microsoft is known for its progressive work on online controlled experiments. In terms of academic research, the fact that the most notable papers are affiliated and funded by corporations, who sell their own marketing solutions can be considered suspicious.

More diversity in terms of authors and affiliations, and contributions from academic institutions would bring even more credibility to A/B testing. That said, the method has been proven to provide substantial value to businesses when applied correctly, and cemented its place in the toolboxes of digital marketers worldwide.

5.1.2 Implications for practice

This thesis has identified common pitfalls in A/B testing that the experimenter should be aware of. The list is not exhaustive, and the findings may not be applicable to every organization. That said, often awareness of the possible ways to fail motivates the experimenter to conduct due diligence on the subject. With A/B testing often being

glorified with radical examples of success stories, it is called for to conduct a critical look into the methodology, and dig into the things that can actually go wrong. The reader of this thesis hopefully walks away with a healthy sense of skepticism towards A/B testing, and knows the critical success factors related to the methodology.

Often the literature on A/B testing is in the form of a practical guide to guide experimenters towards the best practice. One of the more notable papers published on A/B testing is Kohavi's research group's "Controlled experiments on the web:survey and practical guide". Already in 2008, researchers had a good understanding on what are the possible pitfalls when designing and implementing experiments.

The experimenting landscape has changed significantly since 2008 with new technologies and online services allowing new kinds of variables to be tested. The move towards machine learning solutions to conduct experiments is changing the way managers approach experimenting. Taking into account these changes, Kohavi's early findings still hold true, and have been refined further by multiple researchers, such as Bakshy et al. (2014).

5.2 Limitations and future research

This thesis is limited in its ability to review A/B testing as a method and form an exhaustive picture of the methodology and research conducted on it. For the purpose of this thesis, the scope of research was narrowed to include selected databases and keywords based on exploratory searches. The narrower scope allowed for analysis relevant to the online marketing context, but left other fields of science out of the analysis. The findings of this thesis may not apply in other fields, such as medical or social sciences.

Since many of the articles written on A/B testing are still young in academia, many theories are left unrefined. More empirical research conducted on the theories presented in conferences by research groups representing a corporation would be called for in future

research. The growing amount of publications and citations to notable papers indicates that the process is already ongoing.

References

Accenture. 2018. Experimentation is how you do things, *https://www.accenture.com/fi-en/insights/digital/continuous-experimentation, Accessed Oct 16th 2019*

Bakshy, E. & Eckles, D. & Bernstein, M.S. 2014. Designing and Deploying Online Field Experiments*, WWW '14 Proceedings of the 23rd international conference on World wide web, Pages 283-292*

Bakshy, E. & Eckles, D. 2013. *U*ncertainty in online experiments with dependent data: an evaluation of bootstrap methods*, KDD '13 Proceedings of the 19th ACM SIGKDD international conference on Knowledge discovery and data mining, Pages 1303-1311*

Bronzini, A. 2018, The 2018 State of Conversion Optimization Report*, https://conversionxl.com/blog/2018-conversion-optimization-report/, Accessed Oct 15th 2019*

Cruz-Benito, J. & Vázquez-Ingelmo, A. & Sánchez-Prieto, J. et al. 2017. *E*nabling Adaptability in Web Forms Based on User Characteristics Detection Through A/B Testing and Machine Learning*, IEEE Access (Volume: 6)*

Deng, A. & Lu, J. & Chen, S. 2016. Continuous Monitoring of A/B Tests without Pain: Optional Stopping in Bayesian Testing*, 2016 IEEE International Conference on Data Science and Advanced Analytics (DSAA)*

Deng, A. & Shi, X. 2016. Data-Driven Metric Development for Online Controlled Experiments: Seven Lessons Learned*, KDD '16 Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*

Pages 77-86

Deng, A. & Xu, Y. & Kohavi, R. et al. 2013. Improving the sensitivity of online controlled experiments by utilizing pre-experiment data*, WSDM '13 Proceedings of the sixth ACM international conference on Web search and data mining, Pages 123-132*

Dmitriev, P. & Frasca, P. & Gupta, S. et al. 2016. Pitfalls of long-term online controlled experiments*, 2016 IEEE International Conference on Big Data*

Dmitriev, P. & Gupta, S. & Woo Kim, D. et al. 2017. A Dirty Dozen: Twelve Common Metric Interpretation Pitfalls in Online Controlled Experiments*, KDD '17 Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Pages 1427-1436*

Dmitriev, P. & Wu, X. 2016. Measuring Metrics*, CIKM '16 Proceedings of the 25th ACM International on Conference on Information and Knowledge Management*
Pages 429-437

Esteller-Cucala, M. & Fernandez, V. & Villuendas, D. 2019. Experimentation Pitfalls to Avoid in A/B Testing for Online Personalization*, UMAP'19 Adjunct Adjunct Publication of the 27th Conference on User Modeling, Adaptation and Personalization, Pages 153-159*

Fabijan, A. & Dmitriev, P. & Olsson, H.H. et al. 2018. Online Controlled Experimentation at Scale: An Empirical Survey on the Current State of A/B Testing*, 2018 44th Euromicro Conference on Software Engineering and Advanced Applications (SEAA)*

Hauser, J. & Katz, G. 1998. Metrics: you are what you measure!*, European Management Journal, Volume 16, Issue 5, October 1998, Pages 517-528*

Hohnhold, H. & O'Brien, D. & Tang, D. 2015. Focusing on the Long-term: It's Good for Users and Business*, KDD '15 Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. Pages 1849-1858*

Keser, C. (2019) The top 3 mistakes that make your A/B test results invalid*,* *https://www.widerfunnel.com/3-mistakes-invalidate-ab-test-results/ , Accessed Oct 21st 2019*

Kohavi, R. & Deng, A. & Longbotham, R. et al. 2014. Seven rules of thumb for web site experimenters*, KDD '14 Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining, Pages 1857-1866*

Kohavi, R. & Henne, R.M. & Sommerfield, D. 2007. Practical Guide to Controlled Experiments on the Web: Listen to Your Customers not to the HiPPO*, KDD '07 Proceedings of the 13th ACM SIGKDD international conference on Knowledge discovery and data mining, Pages 959-967*

Kohavi, R. & Longbotham, R. 2017. Online Controlled Experiments and A/B Testing. *In: Sammut C., Webb G.I. (eds) Encyclopedia of Machine Learning and Data Mining*

Kohavi, R. & Longbotham, R. 2007. Online Experiments: Lessons Learned*, Computer (Volume: 40 , Issue: 9 , Sept. 2007)*

Kohavi, R. & Longbotham, R. & Sommerfield, D. et al. 2009. Controlled experiments on the web: survey and practical guide*,* Data Min Knowl Disc (2009) 18: 140

Kohavi, R. & Thomke, S.H. 2017. The Surprising Power of Online Experiments: Getting the Most Out of A/B and Other Controlled Tests*, Harvard Business Review 95, no. 5 (September–October 2017), Pages 74–82.*

Linden, G. 2006, Early Amazon: Shopping cart recommendations, *http://glinden.blogspot.com/2006/04/early-amazon-shopping-cart.html, Accessed Oct 15th 2019*

Lu, L. & Chuang, L. 2014. Separation strategies for three pitfalls in A/B testing, *UEO Workshop, KDD. ACM*

Mulrow C. D. 1994. Systematic Reviews: Rationale for systematic reviews*, BMJ 1994, Pages 309 :597*

Optimizely. A/B Testing, *https://www.optimizely.com/optimization-glossary/ab-testing/, Accessed Oct 15th 2019*

Ross, S. 2005, Introductory Statistics*, Elsevier/Academic Press cop. 2005. 2nd ed*

Siroker, D. & Koomen, P. 2013. A/B Testing: The Most Powerful Way to Turn Clicks into Customers, *John Wiley & Sons Incorporated*

Tamburrelli G. & Margara A. 2014. Towards Automated A/B Testing. *In: Le Goues C., Yoo S. (eds) Search-Based Software Engineering. SSBSE 2014. Lecture Notes in Computer Science, vol 8636.*

VWO. What is A/B Testing? (https://vwo.com/ab-testing/#ab-testing-challenges), Accessed Oct 21st 2019

Appendix A: tables

Table 1. Publications by research area for search string "A/B testing" OR "online controlled experiment" in Scopus.

Subject area (Scopus)	**Documents**	**% of 450**
Computer Science	249	55.33%
Engineering	37	8.22%
Decision Sciences	36	8.00%
Mathematics	35	7.78%
Business, Management and Accounting	33	7.33%
Social Sciences	29	6.44%
Medicine	10	2.22%
Arts and Humanities	6	1.33%
Economics, Econometrics and Finance	3	0.67%
Agricultural and Biological Sciences	2	0.44%
Immunology and Microbiology	2	0.44%
Physics and Astronomy	2	0.44%
Psychology	2	0.44%
Chemical Engineering	1	0.22%
Environmental Science	1	0.22%
Health Professions	1	0.22%
Materials Science	1	0.22%

Table 2. Publications by research area for search string "A/B testing" OR "online controlled experiment" in Web of Science

Subject area (Web of Science)	**Documents**	**% of 214**
Computer Science	173	80.84%
Engineering	43	20.09%
Business Economics	30	14.02%
Mathematics	21	9.81%
Educational Research	11	5.14%
Infectious Diseases	9	4.21%
Science Technology	9	4.21%
Immunology	7	3.27%
Information Science	7	3.27%
Pediatrics	7	3.27%
Microbiology	6	2.80%
Pathology	6	2.80%
Communications	5	2.34%
Gastroenterology	5	2.34%
Genetics Heredity	4	1.87%
Health Care Sciences	4	1.87%
Mathematical Computational Biology	4	1.87%
Respiratory System	4	1.87%
Telecommunications	4	1.87%
Automation Control Systems	3	1.40%
Biochemistry	3	1.40%
Medical Informatics	3	1.40%
Public Environmental Occupational Health	3	1.40%
Robotics	3	1.40%
Social Issues	3	1.40%

Table 3. Publications by document type for search string "A/B testing" OR "online controlled experiment" in Web of Science

Document Types (Web of Science)	Documents	% of 214
Conference Paper	159	74.30%
Article	58	27.10%
Other	5	2.34%
Editorial	3	1.40%
Abstract	2	0.93%
Review	2	0.93%
Clinical Trial	1	0.47%

Table 4. Publications by document type for search string "A/B testing" OR "online controlled experiment" in Scopus

Document types (Scopus)	Documents	% of 450
Conference Paper	226	50.22%
Article	58	12.89%
Conference Review	9	2.00%
Book Chapter	3	0.67%
Review	3	0.67%
Undefined	2	0.44%

Table 5. Publications by research area for search string "A/B testing" AND marketing in Scopus.

Research areas (Scopus)	Documents	% of 26
Computer Science	20	76.92%
Business, Management and Accounting	4	15.38%
Social Sciences	3	11.54%
Engineering	2	7.69%
Agricultural and Biological Sciences	1	3.85%
Mathematics	1	3.85%

Table 6. Publications by research area for search string "A/B testing" AND marketing in Web of Science.

Research Areas (Web of Science)	Documents	% of 28
Computer Science	23	82.14%
Business Economics	19	67.86%
Engineering	7	25.00%
Mathematics	6	21.43%
Information Science	2	7.14%
Automation Control Systems	1	3.57%
Communication	1	3.57%
Geology	1	3.57%
Robotics	1	3.57%
Social Issues	1	3.57%
Telecommunications	1	3.57%

Table 7. Publications by document type for search string "A/B testing" AND marketing in Scopus

Document types (Scopus)	**Documents**	**% of 26**
Conference Paper	17	65.38%
Article	9	34.62%

Table 8. Publications by document type for search string "A/B testing" AND marketing in Web of Science

Document Types (Web of Science)	**Documents**	**% of 28**
Conference paper	20	71.43%
Article	9	32.14%
Editorial	1	3.57%

Table 9. Affiliations of documents for search string "A/B testing" OR "online controlled experiment" in Scopus

Affiliation	Documents	%
Microsoft Corporation	31	7.73%
LinkedIn Corporation	14	3.49%
Yandex LLC	13	3.24%
Chalmers University of Technology	11	2.74%
Malmö Högskola	11	2.74%
Facebook, Inc.	9	2.24%
Google LLC	7	1.75%
University of California, Santa Cruz	7	1.75%
University of California, Berkeley	7	1.75%
Outreach.io	6	1.50%
Stanford University	6	1.50%
University of Washington, Seattle	6	1.50%
Carnegie Mellon University	6	1.50%
Leuphana Universität Lüneburg	6	1.50%
AOL Research	5	1.25%
Cornell University	5	1.25%
University of Michigan, Ann Arbor	5	1.25%
Yahoo Research Labs	5	1.25%
eBay, Inc.	5	1.25%